ISSN:0975-9646

Vanishree K et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 233-235

A Study on Association Rules and Clustering Methods

Vanishree K
Department of CSE
PESIT,Bangalore,Karnataka,India

Prakruthi S N
Department of CSE
PESIT, Bangalore,Karnataka,India

Pratiba D
Department of CSE
RVCE,Bangalore,Karnataka,India

Abstract: - Cloud computing relies on sharing of resources to achieve coherence and economies of scale, similar to a utility over a network. At the foundation of cloud computing is the broader concept of converged infrastructure and shared services. Cloud computing has demonstrated that processing very large datasets over commodity clusters can be done by giving the right programming model. Cloud computing is a type of computing that relies on *sharing computing resources* rather than having local servers or personal devices to handle applications. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Cloud computing is a phrase used to describe a variety of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. Mining association rules is one of the most important aspects in data mining. Association rules are dependency rules which predict occurrence of an item based on occurrences of other items. Apriori is the best-known algorithm to mine association rules. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Decision trees are appropriate when there is a target variable for which all records in a cluster should have a similar value. This paper deals about the different mining algorithms applied in cloud environment in order to retrieve required data and also to minimize the computation cost.

Keywords: Cloud, Apriori, Association Rules, Cloud, data-insensitive application, mining algorithms, Decision Tree Algorithm (DTA).

1. INTRODUCTION

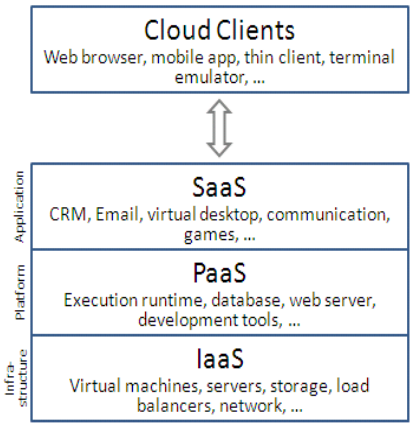
Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps: First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets and the minimum confidence constraint are used to form rules. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Mining Association rule [1] is a way to find interesting associations among large sets of data items. Using this we have determined the frequent item sets based on a predefined support.

By cloud we can say that it is an infrastructure that consists of services delivered through shared data centers and appearing as a single point of access for consumers' computing needs and also provides demanded resources and/or service over the internet. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. Sector automatically replicates files for the better reliability, access and availability [2].

Cloud computing is a phrase used to describe a variety of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet In science, cloud computing is a synonym for distributed computing over a network, and means the ability to run a program or application on many connected computers at the same time [6]. The phrase also more commonly refers to network-based services, which appear to be provided by real server hardware, and are in fact served up by virtual hardware, simulated by software running on one or more real machines [3]. Such virtual servers do not physically exist and can therefore be moved around and scaled up (or down) on the fly without affecting the end user - arguably, rather like a cloud.

2. BACKGROUND AND RELATED WORK

Cloud computing providers offer their services according to several fundamental models: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) where IaaS is the most basic and each higher model abstracts from the details of the lower models. Other key components in anything as a service (XaaS) are described in a comprehensive taxonomy model such as Strategy-as-a-Service, Collaboration-as-a-Service, Business Process-as-a-Service, Database-as-a-Service, etc. In 2012, network as a service (NaaS) and communication as a service (CaaS) were officially included by ITU (International Telecommunication Union) as part of the basic cloud computing models, recognized service categories of a telecommunication-centric cloud ecosystem [4] & [5].



Infrastructure as a service (IaaS)

In the most basic cloud-service model, providers of IaaS offer computers – physical or (more often) virtual machines – and other resources. (A hypervisor, such as Hyper-V or Xen or KVM or VMware ESX/ESXi, runs the virtual machines as guests. Pools of hypervisors within the cloud operational support-system can support large numbers of virtual machines and the ability to scale services up and down according to customers' varying requirements.) IaaS clouds often offer additional resources such as a virtual-machine disk image library, raw (block) and file-based storage, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles. IaaS-cloud providers supply these resources on-demand from their large pools installed in data centers. For wide-area connectivity, customers can use either the Internet or carrier clouds.

To deploy their applications, cloud users install operating-system images and their application software on the cloud infrastructure. In this model, the cloud user patches and maintains the operating systems and the application software. Cloud providers typically bill IaaS services on a utility computing basis: cost reflects the amount of resources allocated and consumed.

Platform as a service (PaaS)

In the PaaS models, cloud providers deliver a "computing platform", typically including operating system, programming language execution environment, database, and web server. Application developers can develop and run their software solutions on a cloud platform without the cost and complexity of buying and managing the underlying hardware and software layers. With some PaaS, the underlying computer and storage resources scale automatically to match application demand so that the cloud user does not have to allocate resources manually.

Software as a service (SaaS)

In the business model using software as a service (SaaS), users are provided access to application software and databases. Cloud providers manage the infrastructure and platforms that run the applications. SaaS is sometimes referred to as "on-demand software" and is usually priced on a pay-per-use basis. SaaS providers generally price applications using a subscription fee.

In the SaaS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Cloud users do not manage the cloud infrastructure and platform where the application runs. This eliminates the need to install and run the application on the cloud user's own computers, which simplifies maintenance and support. Cloud applications are different from other applications in their scalability—which can be achieved by cloning tasks onto multiple virtual machines at run-time to meet changing work demand. Load balancers distribute the work over the set of virtual machines. This process is transparent to the cloud user, who sees only a single access point. To accommodate a large number of cloud users, cloud applications can be multitenant, that is, any machine serves more than one cloud user organization. It is common to refer to special types of cloud-based application software with a similar naming convention: desktop as a service, business process as a service, test environment as a service, communication as a service.

Network as a service (NaaS)

A category of cloud services where the capability provided to the cloud service user is to use network/transport connectivity services and/or inter-cloud network connectivity services. NaaS involves the optimization of resource allocations by considering network and computing resources as a unified whole. Traditional NaaS services include flexible and extended VPN, and bandwidth on demand. NaaS concept materialization also includes the provision of a virtual network service by the owners of the network infrastructure to a third party (VNP – VNO).

3. DESIGN

a. APRIORI ALGORITHM

Apriori is designed to operate on databases containing transactions. Other algorithms are designed for finding association rules in data having no transactions, or having no timestamps (DNA sequencing). Each transaction is seen as a set of items (an itemset). Given a threshold U, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found [7].

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length $k-1$. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of ϵ. Usual set theoretic notation is employed, though note that T is a multiset. C_k is the candidate set for level k. Generate() algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma.count [c] accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero.

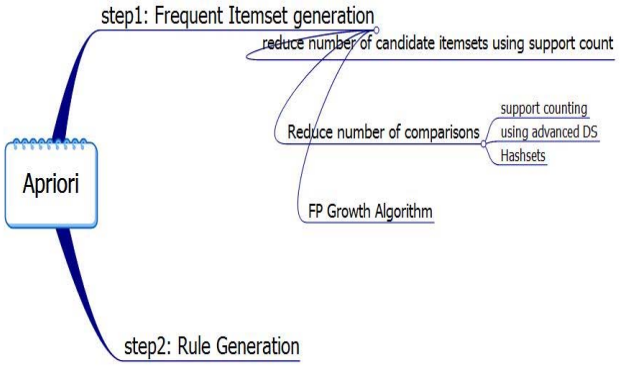


Figure: Apriori Algorithm

$$
\begin{aligned}
&\text{Apriori}(T, \epsilon)\\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\}\\
&\quad k \leftarrow 2\\
&\quad\quad \textbf{while } L_{k-1} \neq \textit{emptyset}\\
&\quad\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\}\\
&\quad\quad\quad \textbf{for transactions } t \in T\\
&\quad\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}\\
&\quad\quad\quad\quad \textbf{for candidates } c \in C_t\\
&\quad\quad\quad\quad\quad \textit{count}[c] \leftarrow \textit{count}[c] + 1\\
&\quad\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \textit{count}[c] \geq \epsilon\}\\
&\quad\quad\quad k \leftarrow k + 1\\
&\quad\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

b. DECISION TREE ALGORITHM

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorisation and generalisation of a given set of data [8]. Data comes in records of the form:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector **x** is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task [9].

PROPOSED SYSTEM

Apriori, while historically significant, suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all $2^{|S|}$-1 of its proper subsets. Later algorithms such as Max-Miner try to identify the maximal frequent item sets without enumerating their subsets, and perform "jumps" in the search space rather than a purely bottom-up approach.

In a decision tree, all paths from the root node to the leaf node proceed by way of conjunction, or *AND*. In a decision graph, it is possible to use disjunctions (ORs) to join two more paths together using Minimum message length (MML). Decision graphs have been further extended to allow for previously unstated new attributes to be learnt dynamically and used at different places within the graph. The more general coding scheme results in better predictive accuracy and log-loss probabilistic scoring. In general, decision graphs infer models with fewer leaves than decision trees.

Alternative search methods

Evolutionary algorithms have been used to avoid local optimal decisions and search the decision tree space with little a priori bias. It is also possible for a tree to be sampled using MCMC in a bayesian paradigm. The tree can be searched for in a bottom-up fashion.

CONCLUSION

In this paper, we proposed a novel clustering technique, called CLTree, which is based on decision trees in classification research. CLTree performs clustering by partitioning the data space into dense and sparse regions. To make the decision tree algorithm work for clustering, we have devised a technique to introduce non-existing points to the data space, and designed a new purity function that looks ahead in determining the best partitioning. CLTree has many advantages over existing clustering methods. Extensive experiments have been conducted with the proposed technique. The results show that it is both effective and efficient.

Also this paper produces a case study on how Decision Tree Algorithm is helpful in retrieving huge data from the complex cloud environment. The DTA (Decision Tree Algorithm) also mine the unknown Data precisely and provide the result as the information collated together. The DTA is not, only algorithm that is used to mine the data but there are other algorithms such as Naive Bayes Algorithm, Clustering algorithm.

REFERENCES

[1] Bhagyashree Ambulkar, Vaishali Borkar, MCA, GHRIIT, MPGINMC – 2012, *Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887*

[2]Minimizing Cost when using Globally Distributed Cloud Services: A Case Study in Analysis of Intrusion Detection Workflow Application Suraj Pandey1, Kapil Kumar Gupta, Adam Barker 2 , Rajkumar Buyya 1, 1 The Cloud Computing and Distributed Systems Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Australia 2 Department of Engineering Science University of Oxford, UK, 7th August 2009.

[3] A Novel Data Mining Algorithm for Semantic Web Based Data Cloud, Kanhaiya Lal, N.C.Mahanti, june 9, 2010, International Journal of Computer Science and Security (IJCSS), Volume (4): Issue (2)

[4] Information Retrieval through Multi-Agent System with Data Mining in Cloud Computing, Vishal Jain,Mahesh Kumar Madan Tech.Appl,Vol 3 (1), 62-66, ISSN:2229-6093.

[5]Data Mining in Cloud Computing, Bhagyashree Ambulkar,Vaishali Borkar, MPGI National Multi Conference 2012 (MPGINMC-2012) 7-8 April, 2012 "Recent Trends in Computing" *Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 - 8887.*

[6] Cloud Computing: Big Data Technology, Michael Farber, Michael Cameron, Joshua Sullivan, 15th , June 2010.

[7] Robert L Grossman, Michael Sabala, Yunhong Gu, Anushka Anand, Matt Handley, Rajmonda Sulo, & Lee Wilkinson. Distributed discovery in e-science: Lessons from the angle project. In Next Generation Data Mining (NGDM '07), 2008.

[8] A Tools-Based Approach to Teaching Data Mining Methods, *Musa J. Jafar, West Texas A&M University, Canyon, TX, USA*, Journal of Information Technology Education: Innovations in Practice, Volume 9, 2010.

[9] J Data Mining In Cloud Computing, Ruxandra – Stefania PETRE, *Database Systems Journal* vol. III, no. 3/2012.